The role of centrality in ambulance dispatching

Seokcheon Lee *

School of Industrial Engineering, Purdue University, 315 N. Grant St., West Lafayette, IN, USA

ARTICLE INFO

Article history:
Received 12 September 2011
Received in revised form 10 May 2012
Accepted 20 May 2012
Available online 26 May 2012

Keywords:
Ambulance dispatching
Centrality
Sensitivity analysis
Weight on centrality
Choice of centrality measure

ABSTRACT

An ambulance dispatching policy, Centrality policy, is proposed in an effort to reduce the response time in demanding emergency situations such as in natural disasters, based on the notion of centrality from the study on complex networks. The nearest neighbor (NN) policy prioritizes the emergency calls by closeness and it has been known effective in the literature. The NN policy is evolved into the Centrality policy by prioritizing the calls based on the centrality in addition to the closeness. The centrality enables to capture the efficiency of a call site in reaching out other current and future calls thus secure the long-term performance beyond the immediate performance pursued by the NN policy. Two parameters are associated with the Centrality policy: weight on centrality and choice of centrality measure. An extensive simulation-based sensitivity analysis is conducted on the algorithmic parameters to examine the role of centrality in ambulance dispatching. The analysis evidences the potential of centrality consideration in reducing the response time beyond the NN policy, given that these parameters are appropriately chosen.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Emergency medical service (EMS) provides pre-hospital treatments to those in need of urgent medical care. The response time in EMS is the time taken to reach the patient after an emergency call is received, and it is of major concern since it might mean the difference between life and death of the patients. Ambulance dispatchers assign appropriate ambulances to the calls such that the response time is minimized. A dispatching decision can be either call-initiated or ambulance-initiated. In call-initiated decisions, a newly arriving call finds idle units (ambulances) thus initiates the decision of selecting a unit among the idle units. On the other hand, if the calls cannot be immediately assigned, they start being queued, and a unit that has just got freed has to choose a call among those waiting thereby initiating the dispatching decision.

The relevance of the two types of dispatching decisions depends on the busyness of the system. Call-initiated decisions are more relevant in routine emergency scenarios where the system load is relatively low, while in high load conditions the ambulance-initiated decisions play the primary role. This research concerns the ambulance-initiated dispatching decisions, in an effort to help respond effectively to the catastrophic natural disasters that recent years have evidenced. According to the Centre for Research on the Epidemiology of Disasters (http://cred.be/), in 2010 alone, 373 natural disasters killed over 296,800 people and affected nearly 208 million others.

One distinct characteristic of the dispatching problem in EMS is that the hospitals serve as hubs of the service as the patients are transferred to hospital. It is also important to recognize the fact that it is not always necessary to transfer the patients to hospital. The actual percentage of essential emergency calls that require transferring to hospital is only 25% in the United States [6]. Therefore, it is highly possible that a unit continues serving multiple calls before heading for a hospital. Various factors would be associated with the probability of transferring to hospital, including the resource scarcity, information uncertainty, crew expertise, and nature of catastrophic event, which essentially differ in space and time.

The objective of this research is to provide an ambulance-initiated dispatching policy in an effort to help ambulance dispatchers to make effective decisions in demanding emergency situations. The dispatching problem under consideration has not been addressed in the literature to the best of the author's knowledge. However, the static version of the problem with a single unit and with the probability of transferring to hospital equal to zero, has been studied under the name of Travelling Repairman Problem (TRP) or Minimum Latency Problem (MLP). The objective of TRP is to find a route that, for a given set of customer locations, minimizes the total response time of customers rather than the total traveling time that is usually pursued by the well-known Vehicle Routing Problems (VRPs). The TRP is known NP-Hard [24], and several researchers have proposed heuristic methods [2,13] or exact methods for special cases [1,12,19,26].

The D-TRP is the dynamic version of the TRP where the requests for service arrive stochastically in random locations. Bertsimas and van Ryzin [5] proposed several dispatching policies for the D-TRP: FCFS (First Come First Served), SQM (Stochastic Queue Median), PART (PARTitioning), TS (Traveling Salesman), SFC (Space Filling...
The Centrality policy has two algorithmic parameters, weight on centrality and choice of centrality measure, which enable the policy to be flexibly applicable to various scenarios. An extensive simulation-based sensitivity analysis is conducted on the algorithmic parameters to examine the role of centrality in ambulance dispatching. The analysis evidences that the centrality consideration, upon the right selection of the parameters, can significantly reduce the average as well as the variation of response time beyond the NN policy which is presumably to be effective across all different scenarios as discussed above. Therefore, the Centrality policy, despite its simplicity, is capable of effectively supporting the decisions of ambulance dispatchers in various demanding emergency situations requiring real-time decision making.

The rest of this paper is organized as follows. Section 2 introduces the Centrality policy. The policy and its parameters are evaluated and analyzed in various scenarios in Sections 3 and 4. Finally, Section 5 concludes this work and discusses future work.

2. The Centrality policy

An ambulance-initiated dispatching policy is devised in this section that is flexibly applicable to various demanding emergency situations. The NN (Nearest Neighbor) policy, dispatching the freed unit to the closest call site, is a policy that can potentially be used in such situations. However, one problematic aspect of the NN policy is that it myopically pursues only the immediate performance without taking into account the long-term consequences. The Centrality policy designed here incorporates the principle of centrality from the study on complex networks, thereby securing the long-term performance.

2.1. The policy

The study of complex networks is an active area of scientific research on large-scale real-world networks. One principal thrust in this area has been the identification of the structural properties that are common to many real networks. Node centrality in a network indicates the importance of a node in the operational efficiency of the network, and various measures of centrality have been defined in an effort to identify the common properties in the distribution of node centrality (Several relevant measures of centrality will be detailed in Section 2.2). The node centrality is used as a decision principle for the dynamic ambulance dispatching problem at hand. When an ambulance gets freed, a network can be constructed where nodes represent waiting calls that have not been assigned to any unit and an edge between every pair of calls has a value of distance between the two call sites connected by the edge. The centrality of a call computed upon this call network can be interpreted as the efficiency of the call in reaching other calls or the density of calls around the call with respect to the geographical call distribution over the service area. When calls are prioritized by the centrality and a unit is dispatched to the most central call, the unit will be given the opportunity, after the completion of the immediate service, to serve the other calls around it at the maximum rate of completion.

The centrality also contributes to the preparedness for future calls. The geographical distribution of current calls reflects the call arrival pattern over the service area. Therefore, when units are positioned according to the centrality (computed upon the network of current calls), they would occupy the regions with high call rates and consequently become able to quickly respond to the calls arriving in the future, thus being well-prepared. The centrality, therefore, implies the efficiency of a call site in serving the current as well as future calls, and the centrality consideration in the dispatching decisions is expected to synergistically escalate the completion rate.

However, if calls are prioritized only by the centrality, the units would travel excessively just to reposition themselves in central nodes without enough exploitation of calls in vicinity. Therefore, it is undesirable to use the centrality alone for the dispatching decision and the centrality has to be combined with a measure that provides the capability of local exploitation. The closeness that is used in the NN policy is an appropriate measure as it enables to pursue minimizing each current response time. The NN policy is evolved into the Centrality policy by prioritizing calls based on the centrality in addition to the closeness, thereby being equipped with both global exploration capability and local exploitation capability. The Centrality policy is presented in four steps as follows.

2.1.1. The Centrality policy

i. When an ambulance gets freed, identify all unassigned calls $U$.

ii. Compute centrality $c_{u}$ of each call $u \in U$ upon the network of calls $U$ with the edge between every pair of calls having the value of distance between them (alternative centrality measures will be detailed in Section 2.2).

iii. Compute goodness $f_{u}$ of each call $u \in U$ based on two quantities: 1) expected (shortest) response time $t_{u}$ to call $u$ and 2) centrality $c_{u}$ of call $u$ weighted by $w (\geq 0)$.

\[
f_{u} = \frac{c_{u}^{w}}{1 + t_{u}}
\]

iv. Dispatch the freed unit to the call $u^{*}$ that maximizes the goodness.

\[
u^{*} = \arg \max_{u \in U} f_{u}
\]

2.2. Algorithmic parameters

Two parameters are associated with the Centrality policy: weight on centrality $w$ (numerical parameter) and centrality measure $c_{u}$ (functional parameter). Note that the Centrality policy is exactly the same as the NN policy when $w = 0$ because then the goodness $f_{u}$ becomes a function of the closeness only; however, when the weight is positive the policy incorporates the centrality into the decision by the extent corresponding to the weight. As mentioned before, one crucial characteristic of the ambulance dispatching is the uncertainty involved in
the need for transferring the patient to hospital. The choice of the weight value will be affected by the probability of transferring as it determines the relevance of the centrality consideration, i.e. the centrality will be more relevant in lower probability. Also, various factors would in some way affect the choice of the weight value, such as the size of service area, size of ambulance fleet, call arrival pattern, etc. The effect of the weight in various operational scenarios will be evaluated and analyzed in Section 3.

The second parameter is the centrality measure \( c_p \). It is computed upon the network of unassigned calls \( U \) with the edge between every pair \( i \) and \( j \) of calls having the value of distance \( \tau_{ij} \) between them. Among various centrality measures available in the literature, three popular centrality measures are chosen as they are appropriate to the call network. The measures are presented as follows and will be evaluated in Section 4.

2.2.1. Weighted degree (WD)

The degree of a node is the number of edges that the node has, and the weighted degree is an extension of the degree to the weighted network (where an edge has a weight representing capacity or strength). The weighted degree of a node is the sum of the weights of all edges connected to the node \([3,20]\). Note that this method is used when higher weight values are preferred (e.g. capacity and strength); however, the weight in the call network represents distance, thus lower weight values are preferred. The weighted degree in this case is computed by the sum of the reciprocals of weights.

\[
k_u = \sum_{i=U,i\neq u} \frac{1}{1+\tau_{ij}}
\]

2.2.2. Distance centrality (DC)

This measure represents the proximity of a node to the rest of nodes in the network \([18,23]\). It is defined as the inverse of farness which is the sum of distances to all other nodes. This measure is also called closeness centrality.

\[
d_u = \frac{1}{1+\sum_{i=U,i\neq u} \tau_{ui}}
\]

2.2.3. Betweenness centrality (BC)

This measure is used to estimate the influence of a node over the flow in a network \([4,11,14]\). It is defined as the sum of fractions of all the shortest paths between each pair of vertices in a network that traverse a given node.

\[
b_u = \sum_{ij \in U} \frac{\sigma_u^{ij}}{\sigma^{ij}}
\]

3. Performance evaluation of Centrality policy

In this section, the performance of the Centrality policy is evaluated in various scenarios implemented in a discrete event simulator, by the performance enhancement over the NN policy as well as another local policy, the DNN policy, that is a more articulate version of the NN policy as will be detailed in Section 3.3. Since the NN policy is provable to be effective across all different scenarios as discussed before, if the performance of the Centrality policy even outperforms the NN policy, the effectiveness of the Centrality policy will get supported to the large extent. In this evaluation, the effect of the weight on centrality is investigated while using weighted degree as centrality measure, and the effect of different centrality measures will be analyzed in Section 4.

3.1. Experimental design

The service area is represented in a 5×5 square grid as shown in Fig. 1. Each vertex generates calls at a certain rate and ambulances move from vertex to vertex through edges each with 1 min of traveling time. Once dispatched to a call, the ambulance serves the patient with a service time that is exponentially distributed with average service time = 0.5 min. The ambulance then, with a probability, transfers the patient to hospital which is located at the center of the grid as depicted in the figure. The probability of transferring to hospital is denoted by hospital_prob\(^1\) throughout the rest of the paper. The purpose of this experiment is to evaluate various ambulance-initiated dispatching policies, and in cases when call-initiated decisions are needed, the policy of dispatching the closest unit is applied as this policy is the most commonly used in practice \([8,10,15]\).

Three factors are taken into consideration in generating different test conditions: 1. call arrival pattern, 2. size of ambulance fleet, 3. hospital_prob. A total of 12,500 calls are generated at the rate of 1 call/min following an exponential distribution, and they are distributed to the vertices according to one of the four call patterns as shown in Fig. 3: a. uniform, b. centered, c. cornered and d. bipartite. A value in the figure of each call pattern represents the probability for an arriving call to be allocated to a corresponding vertex. For example, in the cornered pattern, the vertex located in the center gets an arriving call

\(^1\) According to the statistics collected by United Kingdom Department of Health during May to October 2011 from different regions of United Kingdom, the hospital_prob ranges in 46%–83%. However, note that this probability could deviate from this range in different countries and/or in disaster situations.
with a probability of 0.4. The size of ambulance fleet can be one, three, or five, and the units are randomly located in the beginning of each simulation run. The hospital_prob ranges from 0 to 1 with an increment of 0.1. As a result, 132 test conditions (4 call arrival patterns * 3 sizes of ambulance fleet * 11 hospital_prob) are established. Fifty simulation runs are replicated for each test scenario.

Fig. 2. Effect of weight on performance enhancement by Centrality policy.
the optimal weight value can be searched by a simulation study. As the parameters are becoming revealed over time, and it is recommended to start using a small weight value with the parameters of operating environment are initially unknown to a large extent, in order to maximize the benefit of centrality consideration. In practice, when a disaster breaks out, the parameters are changing rapidly, and it is crucial to make decisions quickly. A more articulate policy keeps going down towards negative improvement (i.e. increase in variation in response time). As discussed before, if the units pursue too much the benefit of centrality, excessive tardy responses can be avoided with the centrality consideration.

3.2. Performance of Centrality policy over NN policy

For each test condition, the Centrality policy is applied with the centrality measure \( c_0 \) fixed to weighted degree (WD) and the weight on centrality \( w \) varying in \( \{0.001, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 2.5, 3\} \). Fig. 2 shows the average reduction in response time by the Centrality policy over the NN policy, i.e. average reduction in response time = (average response time with NN policy — average response time with Centrality policy) / (average response time with NN policy). The Centrality policy dominantly outperforms the NN policy in all different conditions with up to 86% reduction in response time, as long as the weight is not too large. The improvement is significant even with a small weight (\( w = 0.001 \)), and it thereafter tends to increase with higher weight values. Then, after reaching the peak, the improvement keeps going down towards negative improvement (i.e. increase in response time). As discussed before, if the units pursue too much the centrality they will travel excessively just to reposition themselves in central nodes without enough exploitation of calls in vicinity. The nonlinear behavior with the weight on centrality gives rise to the need for carefully choosing the right value of weight according to the operating environment, in order to maximize the benefit of centrality consideration. In practice, when a disaster breaks out, the parameters of operating environment are initially unknown to a large extent, and it is recommended to start using a small weight value with the Centrality policy. As the parameters are becoming revealed over time, the optimal weight value can be searched by a simulation study.

Fig. 3 summarizes the results in Fig. 2 by taking the maximum improvement (from the best weight value) by the Centrality policy for each test condition (Please refer only to the curves indexed by the “Centrality” for now. The “DNN” will be discussed in Section 3.3.). As shown in Fig. 3, the enhancement by centrality exhibits the bell-shaped nonlinearity with the hospital_prob. It is because the increase of hospital_prob amplifies the contribution of centrality consideration due to the high system load, but at the same time it diminishes the contribution as the chance for a unit to continue serving calls before heading for a hospital gets reduced. When the size of fleet increases, the peak point tends to occur at higher hospital_prob. This is because the increase in fleet size reduces the frequency of ambulance-initiated decisions and thus the contribution of centrality becomes eminent in higher load conditions.

The standard deviation of response times in each simulation run is computed to analyze the impact of the centrality consideration on the performance variation. Fig. 4 (please refer to those indexed by the “Centrality”) shows the average reduction in variation by the Centrality policy over the NN policy, i.e. average reduction in variation = (average standard deviation with NN policy — average standard deviation with Centrality policy) / (average standard deviation with NN policy). Note that the weight value used by the Centrality policy in each test condition is the one that produces the maximum improvement in the average response time in that condition, thus being consistent with the weight used in Fig. 3. The overall pattern is similar to the pattern obtained in Fig. 3; however, the reduction in variation is even larger than the reduction in response time. The Centrality policy reduces the variation up to 94%. The reduction in both average and variation implies that excessive tardy responses can be avoided with the centrality consideration.

3.3. Performance of Centrality policy over DNN policy

The NN policy dispatches a freed unit to the closest call site. However, there can be multiple such sites and the NN policy does not specify the action to be taken in that case. A more articulate policy can be formed by letting choose the call site that has the most number
of calls. This policy, called DNN (Densest among Nearest Neighbors) policy, is more specific than the NN policy but it still remains local. Fig. 3 shows the average reduction in response time of the DNN policy over the NN policy, in comparison with the one maximally achieved by the Centrality policy. The DNN policy provides significant advantages against the NN policy with up to 62% reduction in response time. However, the Centrality policy again shows dominant performance even over the DNN policy with up to 30% more reductions, further demonstrating the significance of the centrality consideration in reducing the response time.

Fig. 4 shows the average reduction in variation by the Centrality policy in comparison with the DNN policy, where the Centrality policy uses the same weight values used in Fig. 3. Though the DNN policy effectively reduces the variation up to 85%, the Centrality policy again exhibits dominant performance over the DNN policy with 55% more reductions in variation. From the observations so far, it is possible to argue that the centrality consideration can significantly reduce the average as well as the variation of response time, as a result of equipping with global exploration capability that is lacking in other local policies.

4. Effect of centrality measures

The Centrality policy has the centrality parameter $c_a$ and this section analyzes the effect of different centrality measures. As mentioned before three centrality measures are taken into consideration, which are WD (weighted degree), DC (distance centrality), and BC (betweenness centrality). The experiment made in the previous section is repeated for each centrality measure, but only in the conditions where the Centrality policy produces the most outstanding improvements, i.e. $\text{hospital\_prob} \in \{0.0, 0.1, 0.2\}$ when fleet size $= 1$, $\{0.3, 0.4, 0.5\}$ when fleet size $= 3$, and $\{0.6, 0.7, 0.8\}$ when fleet size $= 5$. 

---

**Fig. 3.** Reduction in response time by Centrality policy in comparison with DNN policy.
Fig. 5 compares the reduction in response time of different centrality measures. The performance of a centrality measure is represented by the maximal improvement over the NN policy from applying different weights on centrality. All the centrality measures make significant improvements in response time, thus being capable of capturing the guidance information for the global exploration. However, note that the WD consistently produces high performance in all different conditions. The maximal improvement is achieved by the WD in 26 cases out of 36 (72%) with slight differences from the best when it is not the best, while the DC makes 6 best cases (17%) and the BC makes 4 best cases (11%). The strength of the weighted degree is also supported by the reduction in variation as shown in Fig. 6. The weighted degree consistently shows superior performance in variation in most cases. The maximal improvement is achieved by the WD in 22 cases out of 36 (61%), while the DC makes 6 best cases (17%) and the BC makes 8 best cases (22%).

Therefore, from these observations, the weighted degree can be considered most suitable as the centrality measure. The weighted degree is also simple to compute thus it is appropriate to the real-time applications. On the other hand, the BC is much more complex to compute since one has to search for all different shortest paths between every pair of nodes, taking a long time especially when the number of calls is very large.

5. Conclusions

A novel ambulance dispatching policy is proposed by the principle of centrality from the study of complex networks. The policy is applicable to various demanding emergency situations such as in disasters requiring real-time decision making. There are two parameters within the policy. One parameter is the weight on centrality. The weight value has to be carefully chosen according to the operating environment, in order to maximize the benefit of centrality consideration. Even a small weight value gives significant benefits; however, the performance gets considerably degraded if the weight is too large. Another parameter is the centrality measure. The weighted degree
among others is recommended as it consistently produces high performance and is computationally simple. The dispatching policy, upon the right selection of the parameters within the policy, can significantly reduce the average as well as the variation of response time, as a result of being equipped with global exploration capability driven by the centrality consideration that is lacking in other local policies.

This work is aligned with recent endeavors that try to apply information technology and decision support systems in disaster management [7,16,17,22,25]. During emergencies, decision making is a challenging task that requires immediate and effective action despite the pressures of incomplete and erroneous information. The policy devised in this research is expected to effectively support the decisions of ambulance dispatchers, when it is implemented by the use of modern computation and communication capabilities available today.

The policy devised here takes into account only the idle unit that has just got freed, despite the possibility that a busy unit can respond more quickly, even after the completion of the currently assigned service, to the call that is otherwise assigned to the idle unit. To further improve the performance by avoiding this inefficiency, all the units need to be taken into account whether they are idle or busy, and it naturally forces to consider all unassigned calls at the same time, leading to an assignment problem that matches between calls and units.

Another important problem in disasters is the emergency commodity logistics problem that involves transporting relief commodities (e.g. food, water, medical aids, etc.) to the affected people. Though the details of the commodity logistics problem are different from the ambulance dispatching problem, they share several key characteristics. Both problems involve vehicle routing and especially aim to minimize response time [21,27,28]. Therefore, the lessons and principles obtained from the ambulance dispatching will provide the basis on which the solution policy for the commodity logistics problem can be established.
Fig. 6. Reduction in variation of different centrality measures.

### References

10. S.F. Dean, Why the closest ambulance cannot be dispatched in an urban emergency medical services system, Prehospital and Disaster Medicine 23 (2) (2008) 161–165.

Seokcheon Lee received the B.S. and M.S. degrees in Industrial Engineering from Seoul National University, South Korea, in 1991 and 1993 respectively, and the Ph.D. degree in Industrial Engineering from Pennsylvania State University, University Park, in 2005. Currently, he is an Assistant Professor in the School of Industrial Engineering at Purdue University. His research interests include decision making techniques from the principles in economics, swarm intelligence, learning, and complex network theory.